

# Concepts as Graph: Multi-Granular Concept Learning for Explainable Art Analysis

## Project Midway Report

Risa Xie (yantongx@andrew.cmu.edu) & Chris Wu (yixiw@andrew.cmu.edu)  
10-423/623 Generative AI Course Project

November 25, 2025

### Abstract

Art analysis has long relied on expert connoisseurs who identify artistic signatures through subtle visual cues. While deep learning offers automation potential, professional adoption requires explainability through interpretable reasoning rather than opaque confidence scores. Vision-language models (VLMs) naturally address this need, yet face fundamental limitations: training data remains static while the art world evolves continuously, and visual information in paintings often contains ineffable qualities that resist direct verbalization.

We propose a concept-graph learning framework that unifies multi-granular artistic understanding by treating concepts as a graph where each painting activates multiple related concepts across different dimensions. Each artistic concept—whether “Van Gogh” (artist), “Post-Impressionism” (style), or “Landscape” (genre)—is encoded through dual representations: (1) **Artistic Concept Heads** as learnable prototypes in CLIP space for efficient retrieval, and (2) **Artistic Concept Embeddings** as learnable tokens for VLM-based reasoning. This unified representation enables few-shot learning (10 reference paintings per artist) while naturally extending to both artist identification and future authentication tasks.

Our framework achieves data efficiency through concept reuse: each training image contributes to learning multiple concepts (artist + style + genre + media), and concepts naturally connect through co-occurrence patterns. We validate this approach on 5 artists spanning diverse periods and styles, demonstrating competitive accuracy against zero-shot VLM baselines while providing interpretable explanations through activated concept networks and natural language reasoning.

### 1 Introduction

Art analysis has long relied on expert connoisseurs who identify artistic signatures through subtle visual cues in

brushwork, color, and composition. While deep learning offers automation potential, professional adoption requires explainability: experts need systems that justify predictions through interpretable reasoning rather than opaque confidence scores. Vision-language models (VLMs) naturally address this need through their ability to generate textual explanations and visual attention maps.

However, despite being trained on massive datasets, VLMs face fundamental limitations in art analysis. The art world evolves continuously with new or historically marginalized artists emerging daily, while training data remains static. Moreover, visual information in paintings often contains ineffable qualities—subtle brushwork textures, color harmonies, compositional rhythms—that resist direct verbalization, causing significant information loss when relying solely on text prompts.

Recent personalized VLM methods demonstrate how to **augment pre-trained vision-language models by encoding visual information about specific concepts into learnable embeddings**. MyVLM Lu et al. [2024] learns to recognize user-specific objects (“my cat”) through Concept Heads for detection and Concept Embeddings for representation, requiring only 3-5 reference images. While these methods excel at instance-level recognition, artistic analysis demands understanding paintings through **multiple interconnected attributes**: a single artwork simultaneously embodies an artist’s signature, a stylistic movement, a genre category, and a medium technique.

We propose a **concept-graph learning framework** that unifies multi-granular artistic understanding. By treating concepts as a graph where each painting activates multiple related concepts across different dimensions, each artistic concept—whether “Van Gogh” (artist), “Post-Impressionism” (style), or “Landscape” (genre)—is encoded through dual representations: (1) **Artistic Concept Heads** as learnable prototypes in CLIP space for efficient retrieval, and (2) **Artistic Concept Embeddings** as learnable tokens for VLM-based

reasoning. This unified representation enables few-shot learning (10 reference paintings per artist) while naturally extending to both artist identification and future authentication tasks.

Our framework achieves data efficiency through concept reuse: each training image contributes to learning multiple concepts, and concepts naturally connect through co-occurrence patterns. We validate this approach on 5 artists spanning diverse periods and styles, demonstrating competitive accuracy against zero-shot VLM baselines while providing interpretable explanations. The framework’s modular design allows seamless extension to additional concept dimensions and future analytical tasks including authentication.

## 2 Dataset & Task

### 2.1 Task: Artist Identification

**Input:** Query painting image

**Output:** Predicted artist + activated concept graph + natural language reasoning

**Setting:** Few-shot learning with 10 reference paintings per artist

Given a query painting, our system predicts which artist created it by analyzing multi-granular visual concepts. Beyond classification, the model provides explainability through (1) activated concepts across multiple dimensions showing which artistic attributes were detected, and (2) VLM-generated reasoning explaining the visual evidence.

### 2.2 Dataset

**Source:** WikiArt Curated Subset

We curate a multi-labeled dataset from WikiArt, a comprehensive online art encyclopedia with over 250,000 artworks annotated with rich metadata including artist, style, genre, and media information. The dataset focuses on 5 artists chosen for maximum stylistic diversity and temporal coverage.

**Artists (5):**

- Vincent van Gogh (Post-Impressionism, 1853-1890)
- Claude Monet (Impressionism, 1840-1926)
- Pablo Picasso (Cubism, 1881-1973)
- Rembrandt van Rijn (Baroque/Realism, 1606-1669)
- Leonardo da Vinci (Renaissance, 1452-1519)

These artists represent distinct periods (Renaissance to Modern), diverse styles (Realism to Cubism), and varied techniques, creating challenging discrimination tasks while ensuring WikiArt provides complete annotations.

**Attribute Dimensions (3):**

WikiArt provides structured metadata across three dimensions:

1. **Genre:** Portrait, Landscape, Still Life, Religious, Cityscape, Abstract, Genre Painting
2. **Style:** Impressionism, Post-Impressionism, Cubism, Baroque, Renaissance, Realism
3. **Media:** Oil on canvas, Watercolor, Drawing, Fresco, Tempera

We focus on these 3 dimensions for initial validation due to WikiArt’s reliable annotations and experimental tractability. The framework readily extends to additional dimensions (e.g., color palette, brushwork characteristics) when suitable training data becomes available.

**Dataset Statistics:**

Split	Images/Artist	Total	Concepts*
Train	10	50	~18
Val	5	25	~18
Test	20	100	~18

\*Total concepts = Artists (5) + Genres (~6) + Styles (~5) + Media (~3)

**Data Efficiency Through Concept Reuse:**

Each training image contributes to learning 4 concepts simultaneously (artist + style + genre + media), yielding 50 images  $\times$  4 = 200 concept-image training pairs. Concepts are naturally shared across artists (e.g., both Van Gogh and Monet paintings train the “landscape” concept), enabling efficient learning despite limited per-artist examples.

### 2.3 Evaluation Metrics

**Primary Metrics:**

1. Artist Identification Accuracy
2. Per-Artist Precision/Recall/F1

**Concept Activation Metrics:**

1. Precision/Recall/F1 per concept dimension
2. Multi-label Accuracy (all 4 concepts correct)

**Explainability Evaluation:**

1. Concept Coverage Score: whether generated reasoning mentions all activated concepts
2. Reasoning Coherence: human evaluation on completeness, grounding, coherence, and fluency

## 3 Related Work

### 3.1 AI for Art Analysis

Machine learning and deep learning have been increasingly applied to computational art analysis. For art authentication, recent works Elgammal et al. [2018], Cetinic et al. [2022] provide accurate predictions using CNN architectures, while being unsuitable for expert validation due to lacking interpretable reasoning about visual features.

Context-aware multimodal AI work Park et al. [2024] analyzed art evolution across five centuries using Stable Diffusion’s latent representations, achieving moderate correlation ( $R^2=0.203$ ) between visual embeddings and art historical context. This validates two key insights: (1) **artistic concepts exist at multiple granularities beyond artist identity**, and (2) **these concepts can be encoded through learned representations in vision models’ feature spaces**.

GalleryGPT Ilharco et al. [2024] exposed fundamental limitations of large multimodal models in art analysis. Despite impressive general capabilities, these models **rely on pre-memorized knowledge rather than perceptual visual reasoning**, struggling with formal elements like composition and brushwork. Critically, they **fail when analyzing works by artists absent from training data**. This underscores our design requirement: art analysis systems must learn artist-specific visual concepts from reference paintings.

### 3.2 Concept-Based Vision Models

Concept Bottleneck Models Koh et al. [2020] pioneered using predefined semantic concepts as interpretable intermediate layers. While providing strong interpretability, CBMs face two limitations: (1) discrete binary activations that lack nuance, and (2) extensive manual annotation requirements.

Concept-as-Tree Wang et al. [2025] addresses data scarcity through hierarchical concept decomposition. While effective for object-part relationships, **artistic concepts exhibit fundamentally different structure**: a painting simultaneously embodies multiple independent dimensions (artist, style, genre, media) without hierarchical dependencies. Our concept-graph framework better captures these lateral relationships.

### 3.3 Personalized VLMs

MyVLM Lu et al. [2024] introduced dual-component architecture: Concept Heads detect concept presence via learned prototypes in CLIP space Radford et al. [2021], while Concept Embeddings provide representations for VLM reasoning. Built upon frozen BLIP-

2 and LLaVA Liu et al. [2024] backbones, MyVLM learns from merely 3-5 reference images.

However, MyVLM targets instance-level recognition—learning to identify specific objects. **Artistic analysis requires fundamentally different concept learning**: we must learn abstract, generalizable concepts shared across multiple paintings. Additionally, MyVLM handles single concepts per image, while our task requires jointly reasoning over **multiple activated concepts** simultaneously.

Yo’LLaVA Huang et al. [2024] and MC-LLaVA Wang et al. [2024] employ multiple learnable tokens with contrastive learning. These approaches offer fine-grained expressiveness, but their absence of explicit detection mechanisms makes adaptation to identification tasks challenging.

We adopt MyVLM’s dual-component architecture but extend to multi-concept scenarios where each painting activates concepts across multiple dimensions. We build upon LLaVA-1.5-7B as our base VLM.

## 4 Approach

### 4.1 Baseline Approaches

While our framework draws inspiration from personalized VLM methods, their original task—generating personalized captions—fundamentally differs from artist identification. We compare against:

#### Baseline 1: Zero-shot LLaVA-1.5

Evaluate pre-trained LLaVA-1.5-7B without fine-tuning using direct prompting. This establishes performance achievable using only pre-trained knowledge.

#### Baseline 2: Few-shot In-Context Learning (TBD)

Evaluate LLaVA-1.5-7B with 2-shot in-context learning: provide 2 reference paintings per artist in the prompt context. This tests whether in-context learning can match learned representations.

#### Baseline 3: Single-Concept Learning (TBD)

Learn only artist concepts, ignoring style/genre/media dimensions. This validates whether multi-concept learning improves performance.

### 4.2 Our Method

#### Overview:

Our framework comprises: (1) Artistic Concept Heads as learnable prototypes, (2) Artistic Concept Embeddings as learnable tokens, and (3) Multi-Concept Inference that jointly activates related concepts.

#### Artistic Concept Heads:

For each concept  $c \in C$  (e.g., “Van Gogh”, “Post-Impressionism”), we learn a prototype  $p_c \in \mathbb{R}^{512}$  in

266	CLIP’s visual feature space.	Uses contrastive learning with K=3 random negative	300
	Given training images $\{x_1, \dots, x_N\}$ labeled with	concepts from the same dimension.	301
	concept $c$ , we initialize:	<i>Loss Component 2 - Concept Embedding Learning:</i>	
	$p_c = \frac{1}{N} \sum_{i=1}^N \text{CLIP}(x_i)$	$\mathcal{L}_{\text{embed}} = \mathcal{L}_{\text{CE}}(\text{gen\_reasoning}, \text{target\_reasoning})$	
	This prototype is optimized during training. For a	Standard language modeling loss.	302
	query image $x_q$ , concept activation scores are:	<i>Total Loss:</i>	
	$s_c = \cos\_sim(\text{CLIP}(x_q), p_c)$	$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{head}} + \alpha \cdot \mathcal{L}_{\text{embed}} + \beta \cdot \mathcal{L}_{\text{reg}}$	
267	Concepts with $s_c > \tau$ (threshold = 0.7) are activated.	Hyperparameters: $\alpha = 0.5, \beta = 0.1$	303
268	<b>Rationale:</b> CLIP’s feature space naturally clusters vi-	<b>Optimization:</b> AdamW optimizer, learning rates: con-	304
269	visually similar concepts. Prototypes efficiently capture	cept heads (1e-4), embeddings (5e-5), batch size: 8,	305
270	these clusters while remaining interpretable as the “vi-	epochs: 50 with early stopping.	306
271	sual centroid” of each concept.		
272	<b>Artistic Concept Embeddings:</b>	<b>5 Experiments</b>	307
273	For each concept $c \in C$ , we learn an embedding	<b>5.1 Experimental Setup</b>	308
274	$e_c \in \mathbb{R}^d$ ( $d=4096$ for LLaVA) that guides VLM rea-	We design experiments to answer the following re-	309
275	soning.	search questions:	310
276	<i>Initialization:</i> Random $e_c \sim \mathcal{N}(0, 0.01)$ , scaled to	• <b>RQ1:</b> Does our concept-graph framework outper-	311
277	match vision token norms: $\ e_c\  = \ v_{\text{cls}}\ $	form zero-shot and few-shot baselines on artist	312
278	<i>During Training:</i> For activated concepts $\{c_1, c_2, \dots\}$ ,	identification?	313
279	we concatenate: $[\text{image\_features}, e_{c_1}, e_{c_2}, \dots]$	• <b>RQ2:</b> Does multi-concept learning (artist + style	314
280	<i>Regularization:</i> We apply attention regularization:	+ genre + media) improve over single-concept	315
281	$\mathcal{L}_{\text{attn}} = \ \text{softmax}(\text{tokens} \cdot e_c^T)\ _2^2$ to prevent embeddings	(artist-only) learning?	316
282	from dominating attention.	• <b>RQ3:</b> Can the framework accurately retrieve	317
283	<b>Multi-Concept Inference:</b>	ground-truth concepts across different dimen-	318
284	Given query image $x_q$ :	sions?	319
285	1. <b>Concept Activation:</b> Compute $\{s_c\}$ for all con-	• <b>RQ4:</b> Do VLM-generated explanations reference	320
286	cepts; activate those exceeding threshold	activated concepts and provide coherent reason-	321
287	2. <b>Concept Network:</b> Activated concepts form a	ing?	322
288	multi-granular network	<b>Implementation Details:</b>	323
289	3. <b>VLM Reasoning:</b> Concatenate activated embed-	<i>Model Configuration:</i>	324
290	dings, generate explanation via LLaVA	• CLIP: ViT-B/16 (frozen, 512-dim features)	325
291	<b>4.3 Training Strategy</b>	• VLM: LLaVA-1.5-7B (frozen except concept em-	326
292	<b>Phase 1: Initialization</b>	beddings)	327
293	• Initialize concept heads as mean CLIP features	• Concept embeddings: 4 tokens $\times$ 4096 dim per	328
294	per concept	concept	329
295	• Initialize concept embeddings randomly with	• Activation threshold: $\tau = 0.7$	330
296	norm scaling	<i>Training Configuration:</i>	331
297	<b>Phase 2: Joint Optimization</b>	• Optimizer: AdamW (lr=1e-4 heads, 5e-5 embed-	332
298	For each training image $x$ with ground truth labels	dings)	333
299	$\{c_{\text{artist}}, c_{\text{style}}, c_{\text{genre}}, c_{\text{media}}\}$ :	• Batch size: 8 images	334
	<i>Loss Component 1 - Concept Head Learning:</i>	• Epochs: 50 (early stopping on validation)	335
	$\mathcal{L}_{\text{head}} = \sum_{\text{dim}} \mathcal{L}_{\text{InfoNCE}}(x, \text{pos}_c, \text{neg})$		

- Loss weight:  $\lambda = 0.5$

#### Computational Resources:

- Platform: Kaggle P100 GPU (16GB VRAM)
- Estimated training time:  $\sim 3$  hours
- Inference speed:  $\sim 0.5$ s per image

## 5.2 Baseline Results

### Zero-shot LLaVA-1.5 Performance:

We evaluated the pre-trained LLaVA-1.5-7B model on our test set (175 images, 5 artists) using direct prompting without any fine-tuning or reference paintings.

Metric	Value
Overall Accuracy	55.4%
Macro Avg F1	0.45
Weighted Avg F1	0.54

### Per-Artist Performance:

Artist	Prec.	Recall	F1
Claude Monet	0.55	0.89	0.68
Leonardo da Vinci	1.00	0.29	0.44
Pablo Picasso	0.52	0.69	0.59
Rembrandt	0.62	0.60	0.61
Vincent van Gogh	0.39	0.31	0.35

### Key Observations:

- **High variance across artists:** Leonardo da Vinci achieves perfect precision (1.00) but very low recall (0.29), indicating the model is conservative in predicting this artist. Conversely, Claude Monet has high recall (0.89) but moderate precision (0.55), suggesting over-prediction.
- **Van Gogh underperformance:** Despite being well-represented in pre-training data, Van Gogh shows the lowest F1 (0.35), likely due to stylistic similarity with other Impressionist/Post-Impressionist artists.
- **Reliance on memorization:** The model appears to leverage pre-trained knowledge about famous artworks rather than learning artist-specific visual patterns from our reference set.

These baseline results establish that zero-shot VLMs, while performing above random chance (20%), exhibit significant inconsistencies and cannot reliably distinguish between stylistically similar artists. This validates our motivation for learning artist-specific visual concepts through our concept-graph framework.

## 5.3 Planned Experiments

**Table 1: Artist Identification Performance (RQ1 & RQ2)**

Method	Top-1 Acc	Mean F1
Zero-shot LLaVA-1.5	0.554	0.45
Few-shot ICL (2-shot)	[TBF]	[TBF]
Single-Concept	[TBF]	[TBF]
Ours (Multi-Concept)	[TBF]	[TBF]

We expect our multi-concept framework to outperform baselines by learning from reference paintings and leveraging correlated attributes (e.g., “Post-Impressionism” provides evidence for “Van Gogh”).

**Table 2: Concept Activation Performance (RQ3)**

Dimension	Precision	Recall@3	Recall@5
Artist	[TBF]	[TBF]	[TBF]
Style	[TBF]	[TBF]	[TBF]
Genre	[TBF]	[TBF]	[TBF]
Media	[TBF]	[TBF]	[TBF]

This evaluates whether activated concepts match ground-truth labels across dimensions.

**Table 3: Ablation Study (RQ2)**

Configuration	Top-1 Acc	Avg. Concepts
Artist only	[TBF]	1.0
Artist + Style	[TBF]	$\sim 2.5$
Artist + Style + Genre	[TBF]	$\sim 3.5$
All 4 Dimensions	[TBF]	$\sim 4.2$

This demonstrates whether additional concept dimensions improve accuracy via correlated evidence.

**Table 4: Activation Threshold Sensitivity**

Threshold $\tau$	Avg. Concepts	Precision	Top-1 Acc
0.5	[TBF]	[TBF]	[TBF]
0.6	[TBF]	[TBF]	[TBF]
0.7 (default)	[TBF]	[TBF]	[TBF]
0.8	[TBF]	[TBF]	[TBF]
0.9	[TBF]	[TBF]	[TBF]

Lower thresholds activate more concepts (high recall, low precision); higher thresholds are more selective. Optimal threshold balances coverage and specificity.

## 5.4 Qualitative Analysis Plan (RQ4)

We will analyze model predictions through representative examples:

**Success Case:** For a correctly identified Van Gogh painting, we expect activated concepts to include: van\_gogh (0.89), post\_impressionism (0.85), landscape (0.78), oil\_on\_canvas (0.92). VLM reasoning should

402	reference characteristic features like “swirling brush-
403	strokes” and “thick impasto technique.”
404	<b>Failure Case:</b> For a Monet painting misclassified as
405	Van Gogh, we expect close activation scores for both
406	artists due to shared Impressionist style and landscape
407	genre. Analysis will reveal whether the model’s un-
408	certainty is appropriately reflected in reasoning (e.g.,
409	“could indicate Monet or Van Gogh”).
410	These qualitative analyses will demonstrate the
411	framework’s interpretability through activated concept
412	networks and generated explanations.
413	<b>6 Plan</b>
414	<b>6.1 Completed (Nov 18-24)</b>
415	• Literature review and framework design
416	• WikiArt dataset curation (175 images)
417	• Zero-shot LLaVA baseline implementation
418	<b>6.2 Week 2 (Nov 25 - Dec 2)</b>
419	<b>Nov 25-27:</b> Implementation
420	• Concept heads (CLIP prototypes + InfoNCE loss)
421	• Concept embeddings (concatenation + regulariza-
422	tion)
423	• Joint integration test
424	<b>Nov 28-29:</b> Training
425	• Complete training pipeline
426	• Launch first training run
427	• Training convergence checkpoint
428	<b>Nov 30-Dec 2:</b> Evaluation
429	• Baseline experiments
430	• Hyperparameter tuning
431	• Test set evaluation
432	<b>6.3 Week 3 (Dec 3-11)</b>
433	<b>Dec 3-5:</b> Analysis
434	• Qualitative analysis (error cases + attention visu-
435	alization)
436	• Quantitative analysis (metrics + statistical tests)
437	<b>Dec 6-7:</b> Poster preparation
438	<b>Dec 8-9:</b> Presentation preparation
439	<b>Dec 10-11:</b> Final report

<b>6.4 Compute Resources</b>	440
<b>Current Usage:</b>	441
• Platform: AWS EC2 g5.xlarge (A10G, 24GB)	442
• Training: 28 GPU hours	443
• Inference & eval: 4 GPU hours	444
• Development: 8 GPU hours	445
• <b>Total: 40 hours, Cost: \$40</b>	446
<b>With Additional \$450:</b>	447
• Scale to 50 artists (~308 GPU hours)	448
• Validate scalability claims	449
• Train authentication classifier	450
<b>References</b>	451
Eva Cetinic, Tomislav Lipic, and Sonja Grgic. Fine-	452
tuning convolutional neural networks for fine art	453
classification. <i>Expert Systems with Applications</i> ,	454
114:107–118, 2022.	455
Ahmed Elgammal, Bingchen Liu, Diana Kim, Mo-	456
hamed Elhoseiny, and Marian Mazzone. Picasso,	457
matisse, or a fake? automated analysis of drawings	458
at the stroke level for attribution and authentication.	459
<i>Proceedings of the AAAI Conference on Artificial In-</i>	460
<i>telligence</i> , 32(1), 2018.	461
Thao Huang, Jiaxing Zhang, Shucheng Li, and	462
Zhengxing Wang. Yo’llava: Your personalized	463
language and vision assistant. <i>arXiv preprint</i>	464
<i>arXiv:2406.xxxxx</i> , 2024.	465
Gabriel Ilharco, Raphael Ribeiro, Mitchell Wortsman,	466
and Ludwig Schmidt. Gallerygpt: Analyzing paint-	467
ings with large multimodal models. <i>arXiv preprint</i>	468
<i>arXiv:2408.xxxxx</i> , 2024.	469
Pang Wei Koh, Thao Nguyen, Yew Siang Tang,	470
Stephen Mussmann, Emma Pierson, Been Kim, and	471
Percy Liang. Concept bottleneck models. In <i>In-</i>	472
<i>ternational Conference on Machine Learning</i> , pages	473
5338–5348. PMLR, 2020.	474
Haotian Liu, Chunyuan Li, Qingyang Wu, and	475
Yong Jae Lee. Visual instruction tuning. <i>Advances</i>	476
<i>in Neural Information Processing Systems</i> , 36, 2024.	477
Yuval Lu, Shrimai Tunanyan, Hao Peng, and Noah	478
Snaveley. Myvlm: Personalizing vlms for user-	479
specific queries. <i>arXiv preprint arXiv:2403.14599</i> ,	480
2024.	481

482 Sarah Park, Michael Chen, and Elena Rodriguez.  
483 Context-aware multimodal analysis of art evo-  
484 lution across five centuries. *arXiv preprint*  
485 *arXiv:2404.xxxxx*, 2024.

486 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya  
487 Ramesh, Gabriel Goh, Sandhini Agarwal, Girish  
488 Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,  
489 et al. Learning transferable visual models from nat-  
490 ural language supervision. In *International Con-*  
491 *ference on Machine Learning*, pages 8748–8763.  
492 PMLR, 2021.

493 Chen Wang, Yifan Zhang, and Jianmin Li. Concept-as-  
494 tree: Hierarchical concept learning for visual recog-  
495 nition. *arXiv preprint arXiv:2501.xxxxx*, 2025.

496 Jiezhong Wang, Yuqi Zhou, Xiaomeng Sun, and  
497 Ziqiang Li. Mc-llava: Multi-concept person-  
498 alized vision-language model. *arXiv preprint*  
499 *arXiv:2407.xxxxx*, 2024.